DOCUMENT RESUME

ED 433 763 HE 032 313

AUTHOR Ronco, Sharron L.
TITLE Deconstructing the Student Assessment of Instruction Instrument: Some Psychometric Issues. AIR 1999 Annual Forum Paper.
PUB DATE 1999-06-00
NOTE 26p.; Paper presented at the Annual Forum of the Association for Institutional Research (39th, Seattle, WA, May 30-June 3, 1999).
PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Factor Analysis; Higher Education; *Psychometrics; *Rating Scales; *Student Evaluation of Teacher Performance; *Teacher Effectiveness; Test Reliability
IDENTIFIERS *AIR Forum; Florida Atlantic University

ABSTRACT

This paper demonstrates the application of common statistical methods to evaluate the dimensionality, reliability, generalizability, and potential biasing factors of the student assessment of instruction (SAI) instrument used at Florida Atlantic University. Findings indicated: (1) factor analysis uncovered just two factors, one describing instruction effectiveness and one describing workload/difficulty; (2) reliability of the class-average response depended on the number of students rating the class, raising questions about use of such averages in tenure and promotion decisions; (3) generalizability analyses indicated the SAIs primarily reflect the effectiveness of the instructor rather than the influence of the course; and (4) positive correlations were found between three potential biasing variables--student ratings, both expected and actual grades, class size, and workload/difficulty. (DB)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 433 763

Deconstructing the Student Assessment of Instruction Instrument:

Some Psychometric Issues

by

Sharron L. Ronco, Ph.D.
Director, Institutional Effectiveness and Analysis

Florida Atlantic University
777 Glades Rd.
Boca Raton, Florida 33431
(561) 297-2665
(561) 297-2590 fax
sronco@fau.edu

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
✓ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
D. Vura
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Paper presented at the 39th Annual Forum of the Association for Institutional Research, Seattle, Washington, May 1999.

BEST COPY AVAILABLE

HE 032 313

AIR
for Management Research, Policy Analysis, and Planning

This paper was presented at the Thirty-Ninth Annual Forum of the Association for Institutional Research held in Seattle, Washington, May 30-June 3, 1999.
This paper was reviewed by the AIR Forum Publications Committee and was judged to be of high quality and of interest to others concerned with the research of higher education. It has therefore been selected to be included in the ERIC Collection of AIR Forum Papers.

Dolores Vura
Editor
Air Forum Publications

Deconstructing the Student Assessment of Instruction Instrument:

Some Psychometric Issues

Abstract

The ubiquitous student assessment of instruction instrument, usually in its Likert Scale, bubble-sheet format, has become an entrenched ritual in higher education classrooms. Despite the vast literature on the psychometric properties of the ratings, concerns persist about the conceptual structure, validity, bias and utility of these instruments, especially if locally constructed. This paper demonstrates the application of common statistical methods to evaluate the dimensionality, reliability, generalizability and potential biasing factors of our student assessment of instruction instrument. The results of the psychometric analysis raised pedagogical concerns about how the instrument should be used.

Deconstructing the Student Assessment of Instruction Instrument: Some Psychometric Issues

Although facing less dire consequences than Socrates, who was executed in 399 BC for corrupting the youth of Athens with his teachings, many faculty worry that they will be unfairly penalized by student evaluations that fail to accurately capture the quality of their teaching. The student assessment of instruction instrument (SAI), usually in its Likert Scale, bubble-sheet format, has become an entrenched ritual in higher education classrooms. Whereas only about 30% of colleges and universities asked students to evaluate professors in 1973, it's hard to find an institution that doesn't today. Nearly 2,000 studies have been completed on the topic of student ratings of instruction, making it the single most studied area of higher education (Wilson, 1998).

The vast research literature on SAI's has focused mainly on the psychometric properties of the ratings, exploring such issues as their reliability, validity, relationship to other variables, and potential biases. Although complete consensus on these issues has not been reached, certain conclusions have been relatively well accepted by researchers and practitioners in the field. Class-average student ratings are: 1) multidimensional, 2) reliable and stable 3) primarily a function of the instructor who teaches the course rather than the course that is taught; 4) relatively valid against a variety of indicators of effective teaching; and 5) relatively unaffected by a variety of variables hypothesized as potential biases (Marsh, 1987). Over the past 25 years, studies support the validity of SAI's over invalidity (Greenwald, 1997). Cashin (1995)

Cynthia Condore assisted with data analysis and produced the data tables for this paper.

concludes that in general, student ratings are more statistically reliable, valid and relatively free from bias or need for control than other data used for evaluation.

With the major validity issues more or less resolved by the large volume of research in the 1970's and 1980's, research on SAI's in the 1990's had dropped to a low level. But in the current climate of assessment and consumer satisfaction, attention to SAI's has resurfaced, especially since these assessments are often the sole measure of an instructor's teaching ability. Faculty claim that student ratings can make or break their careers, despite the limitations of these instruments to accurately measure teaching skills. Studies indicating that SAI's can be compromised by grading leniency, lighter course workloads, and instructor expressiveness, have appeared in recent issues of *American Psychologist* and *Change*. These studies have shaken some long-standing beliefs about the validity of SAI's and led to re-examining some of the assumptions behind a practice which, in many institutions, has gone stale.

The debate comes at a time when our institution is engaged in a review of our student assessment of instruction instrument. The 22-item questionnaire, originally developed by a subcommittee of the University Faculty Council, has completed its pilot year of implementation. There was an agreement to evaluate the instrument itself at the end of the pilot year and recommend any changes before continuing its use. Administration of the SAI is required in all organized classes at all levels. Eight of the 22 items are mandated by the State Board of Regents; for these items, results identified by course, section and instructor must be made publicly available, and are to be used in evaluating faculty instruction. Faculty and administrators are concerned that the new instrument will become part of the evaluation culture without ever having been evaluated itself for reliability and validity.

We began our investigation into the psychometric properties of our internally developed SAI with an examination of some common reliability and validity concerns. The reliability issues covered

here include consistency, or interrater agreement in the SAI, and the generalizability of instructor vs. course effects. Validity concerns focused on the conceptual structure (dimensionality) of the instrument and its discriminant validity, or biasing by variables unrelated to teaching effectiveness. We begin with a look at one of the validity issues, multidimensionality. Understanding the multiple components of the SAI is important to interpreting other psychometric aspects of the instrument.

Multidimensionality

That SAI's, like the teaching they represent, are a multidimensional construct, is supported by common sense as well as a considerable body of empirical research (Marsh & Bailey, 1993). Teachers may be organized but poor communicators, fair in grading but uninspiring, and so on. SAI's that fail to measure the distinct components of teaching are less useful for formative-diagnostic feedback and more likely to suffer from a "halo effect," a generalization from some subjective feeling having nothing to do with effective teaching, that affects responses to all items (Marsh, 1987).

Factor analytic studies have found that well constructed SAI's are usually multidimensional, consisting of several items tapping specific dimensions that experts believe can be judged accurately by students and are important to teaching. They generally focus on aspects of the instructional process, such as preparation of course material, providing feedback and grading, or the products of effective instruction, such as subject-matter expertise, increased interest, positive attitudes toward learning (d'Apollonia & Abrami, 1997). Although the number of factors identified vary, the common ones include:

1. Course organization and planning
2. Clarity, communications skills
3. Teacher-student interaction, rapport
4. Course difficulty, workload
5. Grading and examinations
6. Student self-rated learning

(Centra, 1993).

Instruments with well-defined factor structures include the Students' Evaluation of Educational Quality (SEEQ), Student Instructional Report (SIR), Student Instructional Rating System (SIRS), and IDEA.

Some researchers have found evidence of a 'Global Instructional Factor', usually based on the items measuring overall instructor ratings and overall course ratings, and postulate that students rate specific dimensions of instruction on the basis of their global evaluation (d'Apollonia & Abrami, 1997). Other researchers attribute the existence of global factors to a misinterpretation of factor analysis (Marsh & Roche, 1997).

Our SAI instrument (Figure 1) is divided into two parts, Assessment of Instruction and Course Outcomes. Part I is further subdivided into four sections: Organization and Content, Communication, Interaction with Students, Assignments and Grading, with a section for overall ratings of instructor and course. The existence of the subsections indicates at least the intention to identify different aspects of instructional effectiveness. However, the overall means and standard deviations shown in Figure 1 along with the item correlations detailed in Table 1, indicate that items 1 through 18 may not discriminate well among the dimensions of effective teaching that we had hoped to identify.

Factor analysis provides a test of whether students are able to differentiate among different components of effective teaching and whether the empirical factors confirm the dimensions that the instrument is designed to measure (Marsh, 1987). Following Marsh & Bailey (1993) and Burdsal & Bardo (1986) we applied principal axis factor, a Kaiser normalization, and a direct oblimin (oblique) rotation. Unlike orthogonal rotation, oblique rotation simplifies the factor pattern matrix while allowing for correlation among the factors (Pedhazur & Schmelkin, 1991). The assumption of orthogonality among subscales of an SAI seem particularly inappropriate and contrary to existing theory.

STUDENT ASSESSMENT OF INSTRUCTION

Course Prefix & Number _____ Sequence Number _____ Semester _____ Year _____ Instructor's Name _____

	Mean	Std Dev	EXCELLENT	VERY GOOD	GOOD	FAIR	POOR
PART I – ASSESSMENT OF INSTRUCTION							
ORGANIZATION AND CONTENT							
1. Description of course objectives and assignments	1.79	.54	1	2	3	4	5
2. Use of class time	1.85	.58	1	2	3	4	5
3. Clarity of syllabus	1.83	.56	1	2	3	4	5
4. Overall organization of the course	1.88	.62	1	2	3	4	5
COMMUNICATION							
5. Communication of ideas and information	1.89	.66	1	2	3	4	5
6. Stimulation of interest in the course	1.94	.67	1	2	3	4	5
7. Facilitation of learning	1.94	.64	1	2	3	4	5
8. Clarity of class presentations	1.93	.66	1	2	3	4	5
INTERACTION WITH STUDENTS							
9. Availability to assist students in or out of class	1.79	.57	1	2	3	4	5
10. Respect and concern for students	1.72	.59	1	2	3	4	5
11. Ability to inspire students to put forth their best efforts	1.91	.66	1	2	3	4	5
12. Usefulness of feedback on assignments and exams	1.96	.66	1	2	3	4	5
ASSIGNMENTS AND GRADING							
13. Expression of expectations for performance in this class	1.89	.58	1	2	3	4	5
14. Fairness in grading student work	1.83	.60	1	2	3	4	5
15. Usefulness of text and supplemental learning materials	2.00	.62	1	2	3	4	5
16. Degree to which assignments/exams reflect course objectives	1.83	.57	1	2	3	4	5
OVERALL RATING							
17. Overall rating of instructor	1.78	.64	1	2	3	4	5
18. Overall rating of the course	1.97	.63	1	2	3	4	5
			HIGH	ABOVE AVERAGE	AVERAGE	BELOW AVERAGE	LOW
PART II – COURSE OUTCOMES							
19. Amount learned	1.94	.53	1	2	3	4	5
20. Amount of effort required	1.91	.43	1	2	3	4	5
21. Difficulty of course	2.13	.44	1	2	3	4	5
22. Expected grade	1.66	.41	A	B	C	D	F

FIG. 1. Students Assessment of Instruction Instrument with Overall Mean and Standard Deviations.

BEST COPY AVAILABLE

TABLE 1. Correlations Among Survey Items, All Classes

Item #	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
1	1.00																					
2	.87	1.00																				
3	.91	.83	1.00																			
4	.93	.92	.91	1.00																		
5	.89	.86	.82	.89	1.00																	
6	.86	.83	.78	.84	.93	1.00																
7	.89	.86	.82	.89	.95	.95	1.00															
8	.89	.86	.83	.90	.96	.92	.95	1.00														
9	.80	.75	.74	.77	.81	.81	.83	.81	1.00													
10	.80	.75	.74	.77	.82	.83	.85	.82	.90	1.00												
11	.85	.80	.78	.83	.89	.93	.92	.89	.88	.92	1.00											
12	.86	.80	.79	.83	.86	.86	.88	.86	.86	.86	.91	1.00										
13	.90	.83	.84	.88	.89	.88	.91	.89	.84	.85	.91	.91	1.00									
14	.82	.77	.76	.79	.82	.81	.84	.81	.81	.85	.86	.87	.90	1.00								
15	.83	.77	.78	.81	.82	.80	.84	.82	.75	.76	.81	.80	.84	.80	1.00							
16	.88	.83	.83	.88	.88	.86	.90	.88	.81	.82	.87	.87	.91	.87	.87	1.00						
17	.90	.87	.83	.89	.93	.92	.94	.93	.86	.89	.93	.90	.92	.87	.86	.90	1.00					
18	.89	.85	.82	.88	.90	.91	.92	.90	.80	.81	.89	.86	.89	.83	.85	.91	.92	1.00				
19	.83	.80	.76	.82	.85	.88	.88	.85	.75	.74	.84	.80	.84	.76	.80	.86	.86	.91	1.00			
20	.33	.35	.31	.33	.31	.37	.33	.32	.32	.26	.34	.34	.33	.23	.30	.32	.32	.35	.49	1.00		
21	.24	.29	.23	.25	.21	.26	.22	.22	.23	.17	.24	.22	.23	.16	.21	.22	.23	.25	.37	.85	1.00	
22	.42	.32	.36	.37	.43	.47	.47	.44	.42	.45	.50	.45	.48	49	.43	.48	.44	.49	.49	.08	-.07	1.00

All correlations significant at $p < 0.01$ (2-tailed)
N = 5,308

BEST COPY AVAILABLE

Examination of the factor loadings on Table 2 confirm that SAI items 1 through 19 load heavily on Factor 1. Factor 2 consists of items 20 and 21 only, and item 22, expected grade, loads moderately on Factor 1. Therefore, our SAI measures two factors: one is a general instructional factor (Factor 1) and the other could be called a "workload" factor (Factor 2). The correlations from the factor structure matrix confirm that the factors are not orthogonal.

TABLE 2. Factor Loadings for the Primary Factors with Item-Factor Correlations

	Loadings from Factor Pattern Matrix		Correlations from Factor Structure Matrix	
	Factor 1	Factor 2	Factor 1	Factor 2
Item 1	.93	.03	.94	.28
Item 2	.86	.10	.89	.33
Item 3	.86	.04	.87	.27
Item 4	.91	.05	.93	.30
Item 5	.95	.002	.95	.26
Item 6	.93	.05	.94	.30
Item 7	.97	.004	.97	.26
Item 8	.95	.01	.95	.26
Item 9	.87	.02	.87	.25
Item 10	.91	-.06	.89	.18
Item 11	.95	.01	.95	.27
Item 12	.93	-.004	.93	.24
Item 13	.96	.003	.96	.26
Item 14	.92	-.08	.90	.17
Item 15	.87	.01	.87	.25
Item 16	.94	.01	.94	.26
Item 17	.97	-.0003	.97	.26
Item 18	.94	.04	.95	.29
Item 19	.84	.20	.89	.43
Item 20	.10	.86	.34	.89
Item 21	-.03	.96	.23	.96
Item 22	.53	-.16	.48	-.02

SAI's that lack a well-defined factor structure are often the victims of poorly worded items, or a survey design that encourages halo effect responses or a certain response set. Comparison of our SAI with instruments known to be multidimensional shows that many of our items are vague or ambiguous, instead of clearly describing the instructor behavior that we hope to evaluate (Figure 2). Reversing the scale on some items, such as the example from the IDEA survey on "dry or dull" presentations, helps prevent a fixed response set.

(5) Communication of ideas and information
(6) Stimulation of interest in the course
(7) Facilitation of learning
(8) Clarity of class presentations

Better wording:

Found ways for students to answer their own questions (IDEA)
Promoted teacher-student discussions, as opposed to mere responses to questions (IDEA)
Made presentations that were dry or dull (IDEA)
Summarized material in a manner which aided retention (IDEA)

The instructor seemed genuinely concerned with student progress and was actively helpful (SIR)
My interest in the subject area has been stimulated by this course (SIR)
The instructor summarized or emphasized major points in lectures or discussions (SIR)

The instructor seemed to know when the students did not understand the material (SIRS)

You found the class intellectually challenging and stimulating (SEEQ)
Proposed objectives agreed with those actually taught so you knew where the class was going (SEEQ)

FIG. 2. Our Items. Compared to Wording in Multidimensional Instruments

Reliability

Reliable student ratings are those which measure something consistently. Although reliability can be assessed by computing correlations among responses to different items designed to measure the same component of effective teaching, the reliability of SAI's is probably most appropriately

determined from studies of interrater agreement that assess agreement among different students within the same course (Gillmore et.al., 1978; Marsh, 1987). Reliability varies with the number of raters, with more raters producing higher reliabilities. The reliability of student ratings is not a contested issue. Given a sufficient number of students, the reliability of class-average SAI's compares favorably with that of the best objective tests (Marsh & Dunkin, 1997).

Interrater reliability can be assessed using the intraclass correlation coefficient, an index that compares variation in responses within classes to variations across classes. It is calculated and interpreted within an ANOVA framework, measuring the relative homogeneity of the ratings within the classes under consideration in relation to the total variation among all ratings across all classes. The lower the intraclass correlation coefficient, the larger the variation in responses among the raters within classes, or the lower the variation in average ratings across classes, or both. The larger the intraclass coefficient, the more differentiation there is among classes relative to that among raters within classes (Feldman, 1977; Winer, 1971).

The intraclass correlation coefficients obtained for our SAI by class level and size (Table 3) are very similar to those obtained in other studies (Marsh & Roche, 1997; Cashin, 1995). Researchers agree that reliabilities under .70 should be interpreted cautiously. From the reliability analysis of our SAI , we conclude that some of our items, especially for lower division classes, are particularly unreliable. Item means from classes with fewer than 10 raters, especially lower-division and graduate level courses, should be viewed with caution. Items 20 (Amount of effort required) and 21 (Difficulty of course) would be expected to show lower interrater agreement since they would naturally vary with the preparation and ability that the individual student brings into the class.

There are three sources of error in reliability which are appropriate to mention here. The first underscores the value of constructing clear, unambiguous items that describe specific instructor behaviors. To the extent that survey items are overly broad, ambiguous, or do not allow respondents

to use their full powers of discrimination, the potential for reliability will not be achieved (Doyle, 1975). Likewise, if the student raters lack skill or motivation, or are careless or thoughtless in completing the SAI, reliability may suffer. Research has shown that trained students, even minimally trained ones, produce more reliable ratings. Finally, intraclass reliability assumes that within-class variability is attributable to random error, but there may be patterned differences in ratings linked to different types of students or subgroups in the same class. (Feldman, 1977). The impact of student background, expectations and learning styles on the evaluation of the educational experience is an area where further research is needed.

TABLE 3. Intraclass Correlation Coefficients by Class Level and Size

	Lower Division				Upper Division				Graduate			
	n=5	n=10	n=20	n=40	n=5	n=10	n=20	n=40	n=5	n=10	n=20	n=30
1. Description of objectives & assignments	.25	.43	.79	.88	.46	.78	.82	.90	.59	.59	.85	.93
2. Use of class time	.38	.56	.76	.94	.55	.80	.84	.91	.50	.70	.90	.96
3. Clarity of syllabus	.71	.38	.74	.91	.23	.82	.84	.88	.54	.66	.82	.95
4. Overall organization of course	.54	.55	.79	.93	.45	.85	.86	.92	.54	.71	.90	.96
5. Communication of ideas/information	.60	.67	.82	.95	.52	.81	.86	.91	.65	.76	.89	.95
6. Stimulation of interest in course	.58	.48	.81	.95	.45	.82	.85	.90	.55	.72	.86	.94
7. Facilitation of learning	.43	.59	.81	.94	.31	.79	.84	.91	.68	.74	.87	.95
8. Clarity of class presentations	.61	.58	.82	.95	.59	.82	.85	.92	.70	.78	.86	.94
9. Availability to assist students in/out of class	.45	.53	.76	.90	.31	.81	.77	.84	.48	.62	.90	.94
10. Respect/concern for students	.35	.71	.78	.89	.46	.81	.78	.88	.36	.70	.91	.95
11. Ability to inspire students	.42	.66	.78	.94	.49	.81	.81	.90	.53	.65	.89	.96
12. Usefulness of feedback on assignments/exams	.01	.63	.80	.94	.41	.82	.82	.89	.41	.62	.87	.97
13. Expression of expectations for performance	.40	.68	.80	.93	.48	.82	.82	.91	.57	.58	.88	.96
14. Fairness in grading	.32	.69	.86	.89	.71	.77	.79	.89	.57	.66	.88	.95
15. Usefulness of text/supplemental learning materials	.39	.63	.78	.84	.40	.84	.80	.87	.75	.66	.79	.94
16. Degree to which assignments/exams reflect course	.40	.58	.79	.89	.45	.83	.81	.91	.54	.67	.85	.95
17. Overall rating of instructor	.39	.65	.85	.94	.55	.84	.85	.92	.59	.72	.90	.95
18. Overall rating of course	.45	.65	.80	.92	.41	.80	.84	.90	.56	.72	.88	.96
19. Amount learned	-.12	.44	.68	.88	.37	.77	.79	.91	.55	.59	.79	.93
20. Amount of effort required	-.03	.46	.57	.61	.68	.65	.77	.78	.72	.65	.80	.87
21. Difficulty of course	.13	.50	.65	.64	.54	.57	.75	.82	.76	.68	.82	.84
22. Expected grade	.62	.65	.61	.87	.44	.66	.74	.86	.51	.70	.40	.88

Generalizability

When the results of SAI's are to be used in personnel decisions, it is useful to know whether the evaluations in a particular course or courses represent the instructor's general teaching effectiveness. Generalizability is concerned with how confident we can be that our data accurately reflects the instructor's general teaching effectiveness, not just how effective he or she was in a particular course (Cashin, 1995).

Generalizability research has found that the influence of the instructor who teaches the course is much larger than that of the course that is being taught (Gillmore et. al., 1978; Marsh, 1987; Hogan, 1973). However, we know that some teachers are better suited to teaching certain kinds of courses. Analyses in which the instructor and course effects are confounded do not allow an assessment of the dependability of student rating data for evaluating the instructor alone.

One way to separate the instructor from the course effects is to correlate the average ratings from each of three conditions: two sections of the same course with the same instructor in two successive semesters; two sections of different courses with the same instructor in the same semester; and two sections of the same course with different instructors in the same semester. The first condition provides a measure of interrater reliability for course-instructor combinations of classes, the second condition isolates the instructor effect, and the third addresses the course effect.

Table 4 indicates that there is reasonable agreement in the ratings obtained for the instructor teaching the same course in successive semesters. The correlations for the same instructor-different course are somewhat lower, indicating a substantial fluctuation in student-rated performance of an instructor from course to course. The low correlation for item 15 (rating of text) reflects its association with the course rather than the instructor. The items most generalizable for the instructor are those dealing with stimulation of interest in the course (item 6) and the expected grade (item 22). Correlations for the course effect with instructor effect removed are generally the lowest, indicating

that the nature of the course is of lesser importance for most items. Only those items dealing with course workload and expected grade are moderately correlated for same course-different instructor pairs. The results of our generalizability analyses are almost identical to those obtained by Hogan (1973), and similar to the results of studies by Marsh & Dunkin (1997), Gillmore et. al. (1978) and Marsh (1987).

TABLE 4. Generalizability of Ratings: Correlation between Pairs of Classes

	Same Instructor, Same Course, Two Successive Semesters	Same Instructor, Different Course, Same Semester	Different Instructor Same Course Same Semester
1. Description of objectives & assignments	.60	.38	.24
2. Use of class time	.60	.34	.14
3. Clarity of syllabus	.58	.36	.22
4. Overall organization of course	.61	.39	.19
5. Communication of ideas/information	.68	.41	.16
6. Stimulation of interest in course	.68	.44	.15
7. Facilitation of learning	.68	.41	.16
8. Clarity of class presentations	.67	.38	.14
9. Availability to assist students in/out of class	.63	.34	.18
10. Respect/concern for students	.65	.32	.18
11. Ability to inspire students	.68	.43	.19
12. Usefulness of feedback on assignments/exams	.66	.41	.20
13. Expression of expectations for performance	.64	.42	.20
14. Fairness in grading	.60	.42	.16
15. Usefulness of text/supplemental learning materials	.57	.29	.28
16. Degree to which assignments/exams reflect course	.60	.31	.22
17. Overall rating of instructor	.64	.42	.16
18. Overall rating of course	.63	.38	.21
19. Amount learned	.61	.37	.19
20. Amount of effort required	.63	.32	.32
21. Difficulty of course	.62	.30	.40
22. Expected grade	.78	.44	.49

Biasing Variables

Although research has consistently shown that a wide variety of variables that could potentially influence student ratings actually have little effect, many faculty persist in believing that SAI's are easily biased by a number of circumstances. These often include course difficulty, grading leniency, instructor popularity, class size, and the students' reasons for taking the course (Centra, 1993). Williams & Ceci (1997) resurrected the "Dr. Fox Effect" debate with their case study illustrating the overriding influence of instructor expressiveness on student ratings. In a companion article, Trout (1997) argued that students who want easier courses, fewer assignments, lighter reading and higher grades give higher evaluation scores to faculty who give them what they want. Higher evaluations, in turn, give professors what they want—tenure and promotion. To the extent that SAI's are biased, it is important to understand the impact of the bias and how it can be controlled.

Research into what constitutes bias has been hindered by lack of a consistent definition of bias. One common definition is that SAI's are biased to the extent that they are affected by variables outside the control of the instructor. Marsh (1987) points out that under this definition, grading leniency would not be a bias, since it is clearly the instructor's purview to decide how to grade. He suggests instead that bias should be restricted to variables not related to teaching effectiveness. This muddies the definition for variables like class size, which does affect teaching effectiveness and is therefore not a bias, but unfairly penalizes the instructor for whom class size is not a choice. Thus, we will also investigate variables that are not technically biasing, but whose association with ratings requires control before comparisons across classes can be made.

The extensive research into potential biasing variables has found that several background characteristics are consistently related to student ratings. These are: student motivation and reason for taking the class, expected grades, course level, academic discipline and workload/difficulty. Class size has been found to have some relationship with student ratings. Variables not found to be related

to student ratings include faculty age, gender, race, research productivity; or student age, gender, level or GPA (Cashin, 1995; Marsh, 1987; Centra, 1993; Marsh & Roche, 1997; Marsh & Hocevar, 1991).

The most frequently employed approach to studying the relationship between background influences and student ratings is simply to correlate the class-average student evaluations with the class-average measure of a suspected biasing variable. We chose measures of three potential biasing variables- grade, class size and workload/difficulty, and computed their correlation with SAI items, shown in Table 5.

TABLE 5. Correlation between Items and Suspected Biasing Variables

		Grade	Expected Grade	Class Size	Effort Required	Difficulty of Course
1.	Course Objectives	.28	.42	-.14	.33	.24
2.	Use of class time	.19	.32	-.12	.35	.29
3.	Clarity of syllabus	.24	.36	-.12	.31	.23
4.	Organization of course	.23	.37	-.12	.33	.25
5.	Communication of ideas	.29	.43	-.15	.31	.21
6.	Stimulation of interest	.33	.47	-.19	.37	.26
7.	Facilitation of learning	.32	.47	-.17	.33	.22
8.	Clarity of presentations	.29	.44	-.15	.32	.22
9.	Available to assist students	.31	.42	-.20	.32	.23
10.	Concern for students	.33	.45	-.19	.26	.17
11.	Able to inspire students	.36	.50	-.20	.34	.24
12.	Usefulness of feedback	.32	.45	-.19	.31	.22
13.	Expression of expectations	.34	.48	-.17	.33	.23
14.	Fairness in grading	.34	.49	-.14	.23	.16
15.	Usefulness of text	.28	.43	-.14	.30	.21
16.	Exams reflect objectives	.33	.18	-.17	.32	.22
17.	Rating of instructor	.29	.44	-.16	.32	.23
18.	Rating of course	.33	.49	-.17	.35	.25
19.	Amount learned	.32	.49	-.19	.49	.37
20.	Amount effort required	.05	.08	-.16		.85
21.	Difficulty of course	-.07	-.07	-.10	.85	
22.	Expected grade	.69		-.27	.08	-.07

All Correlations significant at $p < .01$ level (2 tailed)
Some scales were reversed to maintain meaningful direction of association

Two measures of grade were used: the actual class-average grade earned, and the average rating of item 22, Expected grade. Both were positively correlated with all items except Course difficulty, and expected grades were more highly correlated with ratings than actual grades. Most studies of the relation between grades and SAI's have found correlations between .20 and .30 (Marsh & Roche, 1997). There are at least three very different interpretations of this correlation. The grading-leniency hypothesis proposes that instructors who give high grades will be rewarded with high evaluations, a serious bias to SAI's. The leniency interpretation was advocated by researchers who were critical of ratings validity in the 1970's, but diminished later with further construct-validity research (Greenwald & Gillmore, 1997). The teaching-effectiveness (validity) hypothesis proposes that students learn more and earn higher grades from effective teachers, a correlation that is validly reflected in the SAI's. The students-characteristics hypothesis speculates that preexisting student variables, such as prior subject interest, makes for more satisfied students and better learning.

The fact that correlations between expected grades and ratings were higher than for actual grades might lend credence to the grading-leniency hypothesis, in that student expectations were higher than learning, as measured by actual grades. This same finding by Greenwald and Gillmore (1997) was interpreted by them as supportive of the grading-leniency theory, but by other researchers as not disconfirming the teaching-effectiveness hypothesis. Students in whom a teacher strikes a chord will rate the teacher highly and expect their grades to be higher than normal, even though the actual grades may still not be A's (McKeachie, 1997).

Simple correlation coefficients cannot untangle evidence for these competing hypotheses. Path analytic studies have found the strongest support for the students-characteristics and validity hypotheses. Whereas a grading-leniency effect may produce some bias, support for this suggestion is weak and the size of the effect is likely to be unsubstantial (Marsh & Roche, 1997).

Consistent with previous research, our examination of the relation between class size and student ratings showed a weak negative relation, with the size of the relation stronger for items pertaining to the instructor's interactions and interrelationships with students. Class size should not be interpreted as a bias to student ratings; rather, class size does have a moderate effect on some aspects of effective teaching (most instructors teach better in small classes) and these effects are accurately reflected in the SAI. The class size effect clearly illustrates why the multidimensionality of SAI's need to be taken into account, and that ratings on those items particularly influenced by class size should be norm-referenced.

The association between workload (Effort required) and item ratings ranged in the .30's , with the highest correlation (.49) between effort required and amount learned. Course difficulty and item rating showed slightly lower, but still positive, correlations. This contradicts the myth that students will reward easy courses with high evaluations, but confirms substantial previous research that workload/difficulty and ratings are positively correlated, i.e., students give higher ratings in difficult courses where they have to work hard. Since the direction of the workload/difficulty effect is opposite to that predicted as a potential bias, workload/difficulty should not be considered as a biasing variable in SAI's.

In the regression analysis shown in Table 6, all statistically significant predictors accounted for 30% of the variance in item 17, overall rating of the instructor. Expected grade and workload contributed 27% of the variance, with specific disciplines and course levels predicting the rest. Once these variables were taken into account, class size, actual grade in the course, and the semester (fall, spring or summer) were not significant predictors of instructor rating.

TABLE 6. Regression Results for Item 17 (Overall Rating of Instructor)

Model	R	R Square	Adj R Square	Std Error Estimate	Change Stat R Square Change	F Change	df1	df2	Sig F Change
1[a]	.44	.19	.19	.57	.19	1199.1	1	5048	.0000
2[b]	.52	.27	.27	.54	.08	570.8	1	5047	.0000
3[c]	.54	.29	.29	.54	.01	82.6	1	5046	.0000
4[d]	.54	.29	.29	.54	.01	48.6	1	5045	.0000
5[e]	.55	.30	.30	.53	.01	33.5	1	5044	.0000
6[f]	.55	.30	.30	.53	.002	11.3	1	5043	.0008
7[g]	.55	.30	.30	.53	.002	12.1	1	5042	.0005
8[h]	.55	.30	.30	.53	.002	13.7	1	5041	.0002

a. Predictors (constant), Expected grade
b. Predictors (constant), Expected grade, Effort required
c. Predictors (constant), Expected grade, Effort required, Arts & Letters
d. Predictors (constant), Expected grade, Effort required, Arts & Letters, Graduate
e. Predictors (constant), Expected grade, Effort required, Arts & Letters, Graduate, Education
f. Predictors (constant), Expected grade, Effort required, Arts & Letters, Graduate, Education, Liberal Arts
g. Predictors (constant), Expected grade, Effort required, Arts & Letters, Graduate, Education, Liberal Arts, Upper Level
h. Predictors (constant), Expected grade, Effort required, Arts & Letters, Graduate, Education, Liberal Arts, Upper Level, Urban & Public

Summary and Conclusions

The factor analysis of our SAI uncovered evidence of just two factors, one describing instructional effectiveness and one describing workload/difficulty. Since it is generally agreed that good teaching is multifaceted, it makes sense to measure students' perceptions of teaching as accurately as possible, then treat such measurements as different aspects of teaching. One way to improve the dimensionality of our SAI might be to make the items more descriptive of visible behavioral attributes of teachers. Items whose meanings require less inference from students will result in less ambiguous ratings and guard against the halo effect of having students rate all dimensions on the basis of some general feeling about the class. For instructors who use ratings for improvement, specific behavioral items are more helpful than general overall ratings. Faculty and

administrators who rely on single omnibus items such as "Overall rating of instructor" have an overly simplistic representation of students' perceptions of the instructor. In addition, single items can be unreliable and their use is risky and questionable especially when used for tenure, promotion or salary considerations (Burdsal & Bardo, 1986).

The reliability of the class-average response depends on the number of students rating the class. Item means from classes with fewer than 10 raters, especially lower-division and graduate level courses, should be interpreted cautiously. For tenure and promotion decisions, the number of students rating the course and the number of courses rated should be considered. By providing standard deviations along with item means, we can encourage faculty and administrators to pay as much attention to the variability in responses as they do to the average response. The percent of students enrolled in the class who fill out class evaluations should also be reported, as it can affect the context for reliability. Generally, if two-thirds or more of the students in a class respond, the results can be considered representative of the class as a whole (Centra, 1993). Reliability can also be improved by using unambiguous survey items and by training students to be thoughtful evaluators.

The results of the generalizability analyses indicate that our SAI's primarily reflect the effectiveness of the instructor rather than the influence of the course. Since there is some course effect, however, ratings for a given instructor should be measured across different courses to enhance generalizability. Gillmore et. al. (1978) suggest that when making determinations about an instructor's teaching effectiveness, at least five different courses should be used. If there are fewer than 10 students in any of the classes, data from additional classes are recommended. If a longitudinal archive of student ratings is maintained, a systematic examination of whether teaching quality might be improved by careful assignment of teachers to courses can be made. Finally, since the course does not appear to be a major factor in the determination of SAI ratings, these ratings should not be used to make decisions about a course across instructors.

Our investigation into potential biasing variables on the SAI focused on three areas; grades, both expected and actual, class size and workload/difficulty. Positive correlations between grades and ratings can signal a grading leniency bias, where students reward easy-grading teachers with high evaluations, but this is by no means the only inference to be made. Better teachers produce better learning, and students who come to the class predisposed to do well, either by interest or ability, probably will. The grading-rating correlation can also be explained if the instructor teaches to the better students, which most do. A correlational approach in inadequate to demonstrate bias. Studies using experimental manipulation of the suspected biasing variable have been attempted, but are difficult to validate since they may rely on overly-contrived or unethical manipulation, such as deliberately leading students to believe they have earned higher grades than they have. The effect of bias can also be examined by comparing other indicators of effective teaching with SAI's, such as instructor self-evaluations or ratings by colleagues or administrators.

Class size is not technically a biasing variable since it is does have a moderate effect on effective teaching. Because of the association between class size and ratings, however, comparisons of instructor ratings should not be made across different class sizes. The positive correlations between workload/difficulty and ratings signal that students like to be engaged and challenged, and argue against the hypothesized workload/difficulty bias of SAI's.

Marsh (1987) points out that there are nearly an infinite number of variables that could be related to student ratings and could be posited as potential biases, and that the search for bias has in itself been so biased that it could be called a witch hunt. Methodological shortcomings, a misunderstanding of the definition of bias, and simplistic arguments that correlation indicates bias have plagued the study of potential biasing variables to student ratings, are an injustice to the field, and must not be tolerated.

Limitations and Further Research

This investigation into the psychometric properties of our SAI's has been limited to specific reliability and validity issues using common methodologies, and does not encompass all of the possible approaches to evaluation of these instruments. SAI's are difficult to construct-validate because there is no single criterion of effective teaching. Students' learning is the most widely accepted criteria of effective teaching, but there are others, including the ratings of former students, student achievement in multisection validity studies, faculty self-evaluation of their own teaching effectiveness, and observations of trained observers (Marsh & Roche, 1997; Marsh, 1997). Narrowly defining teaching effectiveness as student learning on a final exam has resulted in studies like Williams and Ceci's (1997) which disparaged the validity of SAI's through their demonstration that an enthusiastic instructor can elicit higher evaluations but not higher test scores. As W.M. Plater pointed out in a response to that case study, the fact that students under the influence of a more enthusiastic teacher think they learn more should be encouraging. Students whose interest in learning is increased by an enthusiastic teacher may not demonstrate mastery in that content until several semesters or even years have passed. There has been little research into other important student outcomes such as motivation, study strategies, future course selection and career aspirations.

We continue to know little about how students arrive at their judgements of teaching effectiveness. British researchers have found a strong correlation between students' approaches to learning and their preferences for teaching styles. "Surface learners" may be satisfied with lectures, while "deep learners" prefer a more facilitative approach. The same students may have different approaches in different classes, depending on their interest in the subject matter (Entwistle & Tait, 1995). There has been little research into how these learning orientations affect SAI's.

Despite the continuing debate over student ratings, the fact remains that there is no readily available alternative method of evaluating instruction . SAI's are often the only vehicle available to

students for communicating their learning needs. Faculty do find them useful, both for formative and summative purposes, and are willing to use them for instructional improvement (Schmelkin et. al., 1997). In general, experts recommend that student ratings form only one component of comprehensive systems of faculty evaluation, and that all indicators of teaching effectiveness undergo the same rigorous examination for validity and reliability to which we have subjected the SAI's.

Marsh (1997) claims that homemade SAI's are rarely evaluated in relation to rigorous psychometric considerations and revised accordingly. Our investigation into the psychometric properties of our SAI addresses the need for such evaluation and revision. As a result of this study, the University Faculty Council voted to scrap the current SAI and revise the instrument, the processes for its administration, and its use in faculty evaluation.

References

Burdsal, C.A. and Bardo, J.W. (1986). Measuring students' perceptions of teaching: Dimensions of evaluation. *Educational and Psychological Measurement* 46: 63-79.

Cashin, W.E. (1995). Student ratings of teaching: The research revisited. IDEA paper No. 32, Center for Faculty Evaluation and Development, Kansas State University.

Centra, J.A. (1993). *Reflective Faculty Evaluation.* San Francisco: Jossey-Bass Publishers.

D'Apollonia, S. and Abrami, P.C. (1997). Navigating student ratings of instruction. *American Psychologist* 52(11): 1198-1208.

Doyle, D.O. (1975). *Student Evaluation of Instruction.* Lexington, MA: Lexington Books.

Entwistle, N. and Tait, H. (1995). Approaches to studying and perceptions of the learning environment across disciplines. *New Directions for Teaching and Learning* 64: 93-103.

Feldman, K. A. (1977). Consistency and variability among college students in rating their teachers and courses: A review and analysis. *Research in Higher Education* 6: 223-274.

Gillmore, G.M., Kane, M.T. and Naccarato, R.W. (1978). The generalizability of student ratings of instruction: Estimation of the teacher and course components. *Journal of Educational Measurement* 15(1): 1-13.

Greenwald, A.G. (1997). Validity concerns and usefulness of student ratings of instruction. *American Psychologist* 52(11): 1182-1186.

Greenwald, A.G. and Gillmore, G.M. (1997). Grading leniency is a removable contaminant of student ratings. *American Psychologist* 52(11): 1209-1217.

Hogan, T.P. (1973). Similarity of student ratings across instructors, courses and times. *Research in Higher Education* 1: 149-154.

Marsh, H.W. (1983). Multidimensional ratings of teaching effectiveness by students from different academic settings and their relation to student/course/instructor characteristics. *Journal of Educational Psychology* 75(1): 150-166.

Marsh, H.W. (1987). Students' evaluations of university teaching: Research findings, methodological issues, and directions for future research. *International Journal of Educational Research* 11: 253-388.

Marsh, H.W. and Bailey, M. (1993). Multidimensional students' evaluations of teaching effectiveness: A profile analysis. *Journal of Higher Education* 64(1): 1-18.

Marsh, H.W. and Dunkin, M.J. (1997). Students' evaluations of university teaching: A multidimensional perspective. In R.P. Perry and J.C. Smart (Eds.), *Effective Teaching in Higher Education: Research and Practice* (pp. 241-320). New York: Agathon Press.

Marsh, H.W. and Hocevar, D. (1991). The multidimensionality of students' evaluation of teaching effectiveness: The generality of factor structures across academic discipline, instructor level and course level. *Teaching and Teacher Education* 7(1): 9-18.

Marsh, H.W. and Roche, L.A. (1997). Making students' evaluations of teaching effectiveness effective. *American Psychologist* 52(11): 1187-1197.

McKeachie, W.J. (1997). Student ratings: Validity of use. *American Psychologist* 52(11): 1218-1225.

Pedhazur, E.J., and Schmelkin, L.P. (1991). *Measurement, Design and Analysis: An Integrated Approach*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Schmelkin, L.P., Spencer, K.J. and Gellman, E.S. (1997). Faculty perspectives on course and teacher evaluations. *Research in Higher Education* 38: 575-592.

Shrout, P.E. & Fleiss, J.L. (1979). Intra-class correlation: Uses in assessing rater reliability. *Psychological Bulletin* 86(2): 420-428.

Trout, P.A. (1997). What the numbers mean. *Change* 29(5): 25-30.

Williams, W.M. and Ceci, S.J. (1997). "How'm I doing?" Problems with student ratings of instructors and courses. *Change* 29(5): 12-23.

Wilson, R. (1998). New research casts doubt on value of student evaluations of professors. *The Chronicle of Higher Education*, January 16, 1998, A12-A14.

Winer, B.J. (1971). *Statistical Principles in Experimental Design (2nd edition)*. New York: McGraw-Hill.